Development of a Text-Based Classifier Tool to Identify Immunology-Related NIH Grants

Fenglou Mao, Charles Hackett, Daniel Rotrosen, Dawei Lin

Division of Allergy, Immunology and Transplantation (DAIT), National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Bethesda, MD Email: fenglou.mao@nih.gov

Abstract

The abstracts of all NIH-funded grants are publically available via the NIH RePORT Expenditures and Results Tool (RePORTER) web site and are a valuable source to study trends in immunology research. To specifically identify immunology-related grants, we have developed a text-based classifier tool, which is capable of analyzing a large number of abstracts and can provide an objective numeric measure of the relevancy of each grant to immunology research.

The classifier tool was developed based on a Bayesian algorithm and more than 1100 grants funded from 2008 to 2013 after review by the NIH Immunology (IMM) Integrated Review Group (IRG), which is the major IRG for immunology-related applications. This set of grants served as the true positive data set. The grant set of true negative data contained more than 1900 NIH grants, funded after review by IRGs whose specific research focus was not immunology related. The classifier tool developed two metrics, Edge-Ratio and Average-Weight, with high discriminative power based on LIKE-score, a similarity score between all collected grants using their "Research, Condition, and Disease Categorization (RCDC)" signatures. RCDC is a computerized reporting process used by NIH since 2008 to categorize research topics for congressionally-mandated report purposes. For a given grant, Edge-Ratio is defined as the percentage of non-zero LIKE-scores between this grant and all grants in the positive set. Average-Weight is defined as the average non-zero LIKE-score between this grant and all grants in the positive set. A Naïve-Bayesian classifier trained by these two features had an overall prediction accuracy of 92.3%. A test of IMM grants funded from 2014 to 2015, which were not included in the training sets, yielded a value of 80.1% accuracy.

Grant Training Set Construction

Fiscal Year	Number of IMM Grants	Number of Non-IMM Grants
2008	125	343
2009	164	332
2010	121	311
2011	115	264
2012	129	297
2013	476	353
Total	1130	1900

RCDC LIKE-Score Calculation

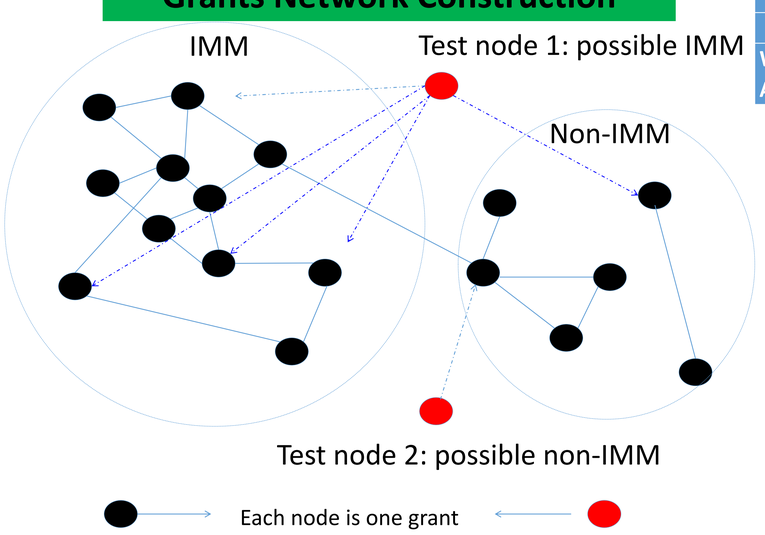
RCDC Terms	Base Publication Weight	Scale Factor	Base Publication Weight(Scaled)	Matching Publication Weight	Dot Product
Basophils	79	3.03	239	212	507
Bone Marrow	93	1.59	148	122	180
Boxing	354	2.05	726		0
Cations	34	2.21	75	123	92
Cell Line	750	1.19	893	123	1098
Cell Nucleus	559	1.49	833	321	2674
Cells	524	0.55	288		0
Chromosomes	222	1.74	386	456	1761
Complement	829	1.33	1103	345	3804
Eicosanoids	186	2.72	506	345	1745
				Like Score	11862

Base Publication Weight(Scaled)=Base Publication Weight*Scale Factor

Dot Product=Base Publication Weight(Scaled)*Matching Publication Weight/100

Like Score=Sum(Dot Product)

Grants Network Construction



Features Extraction

Given an award group G (IMM, et al), and another award A, the average LIKE-score and link ration between A and G are defined as below:

Average LIKE-score=sum(S)/N

Link ratio=N/N_0

S: non-zero LIKE-score between A and any award in G

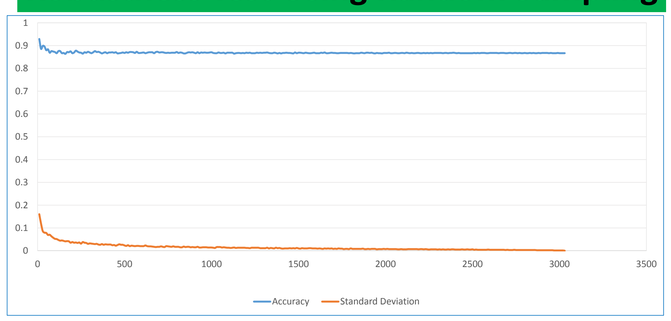
N: number of non-zero LIKE-score between A and any award in G

N_0: number of awards in G

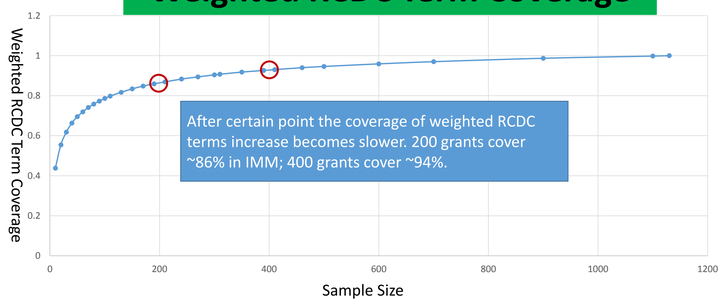
Naïve Bayesian Classifier Performance

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.867	0.043	0.923	0.867	0.894	0.835	0.973	0.950	IMM
	0.957	0.133	0.924	0.957	0.940	0.835	0.973	0.985	Non_IMM
Weighted Avg.	0.923	0.099	0.923	0.923	0.923	0.835	0.973	0.972	

Robustness Test Using Random Sampling



Weighted RCDC Term Coverage



Top 60 Weighted RCDC Terms

T-Lymphocyte	Mediating	T-Cell Development	Testing
B-Lymphocytes	CD4 Positive T Lymphocytes	Inflammatory	Complex
Antigens	Immune	Production	T memory cell
Immune response	receptor	Ligands	Human
Cells	response	Binding	Antigen-Presenting Cells
Mus	T-Cell Receptor	Proteins	Natural Immunity
Role	Immune system	Memory	Natural Killer Cells
Autoimmunity	Molecular	Development	Cell physiology
Regulation	in vivo	Genes	Receptor Signaling
Signal Transduction	Lymphocyte	Toll-like receptors	Virus Diseases
cytokine	Regulatory T-Lymphocyte	T cell response	thymocyte
Immunity	pathogen	Autoimmune Process	Antibodies
Autoimmune Diseases	T-Cell Activation	Inflammation	transcription factor
Dendritic Cells	Infection	Play	Interferons
CD8B1 gene	Pathway interactions	macrophage	novel

Robustness Testing Using Independent and Longitudinal Data

Training Set FY	IMM In Training Set	Non_IMM In Training Set	Accuracy In Training Set	Test Set FY	IMM In Test Set	Non_IMM In Test Set	Accuracy In Test Set
2008	125	343	80.8%,94.2%, 90.6%	2009-2013	1005	1557	78.3%,95.9%,89.0%
2008-2009	289	675	84.1%,94.1%,91.2%	2010-2013	841	1225	83.1%,96.5%,91.0%
2008-2010	410	986	85.1%,95.4%,92.4%	2011-2013	720	914	83.3%,96.3%,90.6%
2008-2011	525	1250	86.3%,95.9%,93.1%	2012-2013	605	650	85.5%,95.8%,90.8%
2008-2012	654	1547	86.5%,95.9%,93.1%	2013	476	353	82.6%,96.9%,88.7%
2008-2013	1130	1900	86.7%,95.7%,92.3%	2014-2015	166	3611	80.1%,93.4%,92.8%

Comparisons with Human Expert Assessments

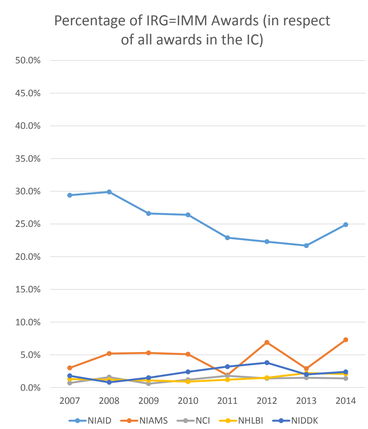
IRG	Total	Predicted	Predicted Number	IRG Reassignment	Strongly Agree	Weakly Agree	Hard To Decide	Total	Ratio
IMM	158	IMM	70	IMM	56	0	0	56	80%
				Non_IMM	11	2	0	14	20%
		Non_IMM	88	IMM	47	8	1	56	64%
				Non_IMM	30	1	1	32	36%
Non_IMM	131	IMM	79	IMM	23	1	2	26	33%
				Non_IMM	44	7	2	53	67%
		Non_IMM	52	IMM	0	0	0	0	0%
				Non_IMM	52	0	0	52	100%

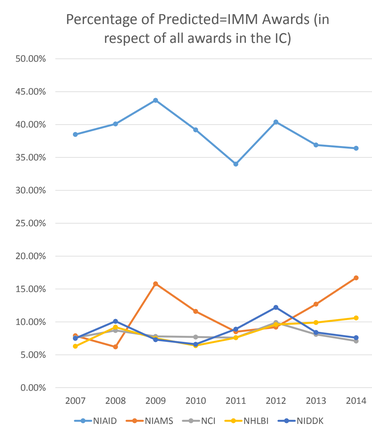
Classification on Other IRGs' Type 1 R01 Awards From 2008-2014

	Predicted As		
IRG	IMM	Non_IMM	Percentage
IMM	676	179	79.06%
OTC	151	1071	12.36%
IDM	142	782	15.37%
AARR	142	787	15.29%
CVRS	115	980	10.50%
DKUS	97	592	14.08%
CB	87	1033	7.77%
MOSS	84	829	9.20%

	Predicted		
IRG	IMM	Non_IMM	Percentage
BDCN	83	1273	6.12%
VH	66	662	9.07%
OBT	51	985	4.92%
BST	47	740	5.97%
EMNR	47	823	5.40%
SBIB	41	974	4.04%
MDCN	33	1063	3.01%
BCMB	28	949	2.87%

Classification on R01 Applications of 5 NIH ICs from Fiscal Year 2007 to 2014





Conclusion

In this study, we have derived a stable text classifier for NIH Immunology Integrated Research Group (IRG), in which majority of immunology research grant applications are reviewed and later funded. The classifier uses a Naïve Bayesian model trained by two network-based features that can be easily calculated using the network of grants constructed using RCDC profile similarity. Various robustness tests including cross validation among independent and longitudinal datasets as well as comparisons with the assessments by human experts have demonstrated that the classifier can provide reasonable prediction performance. The classifier can be used as an objective and quantitative measure for trends and grant portfolio analysis.

Acknowledgement

We appreciate the stimulating discussions from NIAID/DAIT colleagues - Thomas Esch, Lara Miller, Wolfgang Leitner, Travis Hauguel, Ashley Xia, Quan Chen and Susan Cooper, from NIH/OD/QVR team - Brian Haugen and Donald Tiedemann, and computing support from NIAID/OCICB, and grant access help from NIAID/DEA – Virginia Pool.